Exponential Capacity in an Autoencoder Neural Network with a Hidden Layer

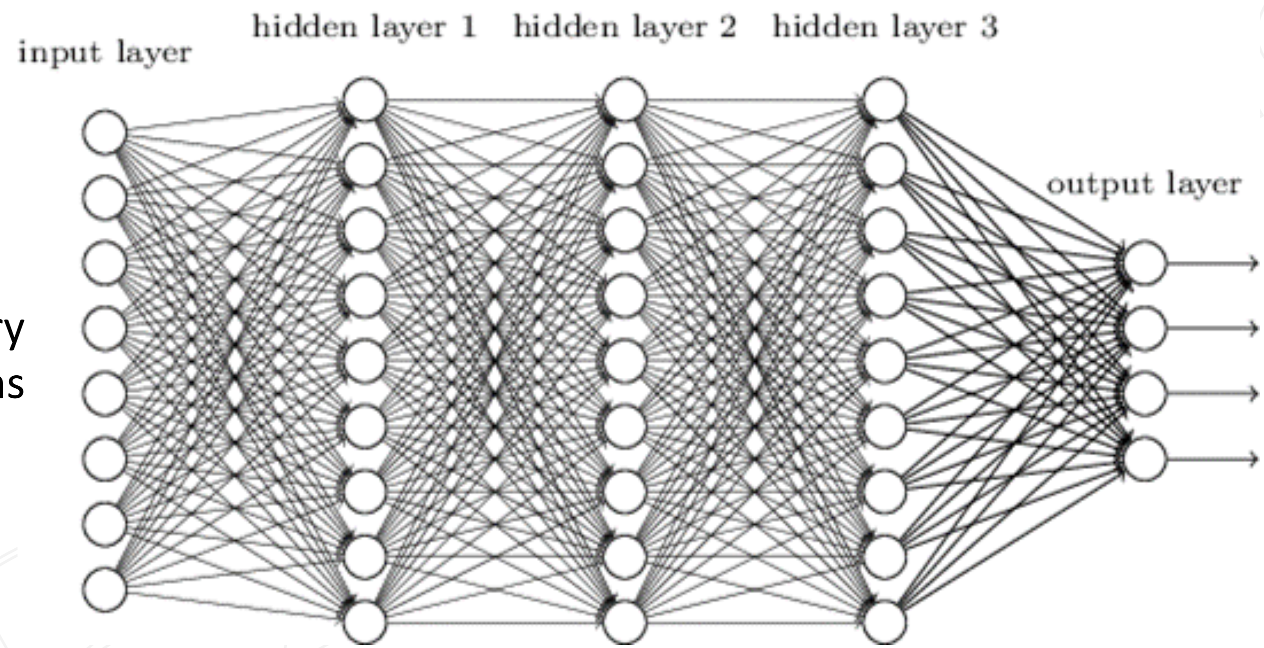
Alireza Alemi*, Alia Abbara

PHYSICS INFORMED MACHINE LEARNING

Santa Fe, NM, Jan 2018

Limits of computations in deep nets?

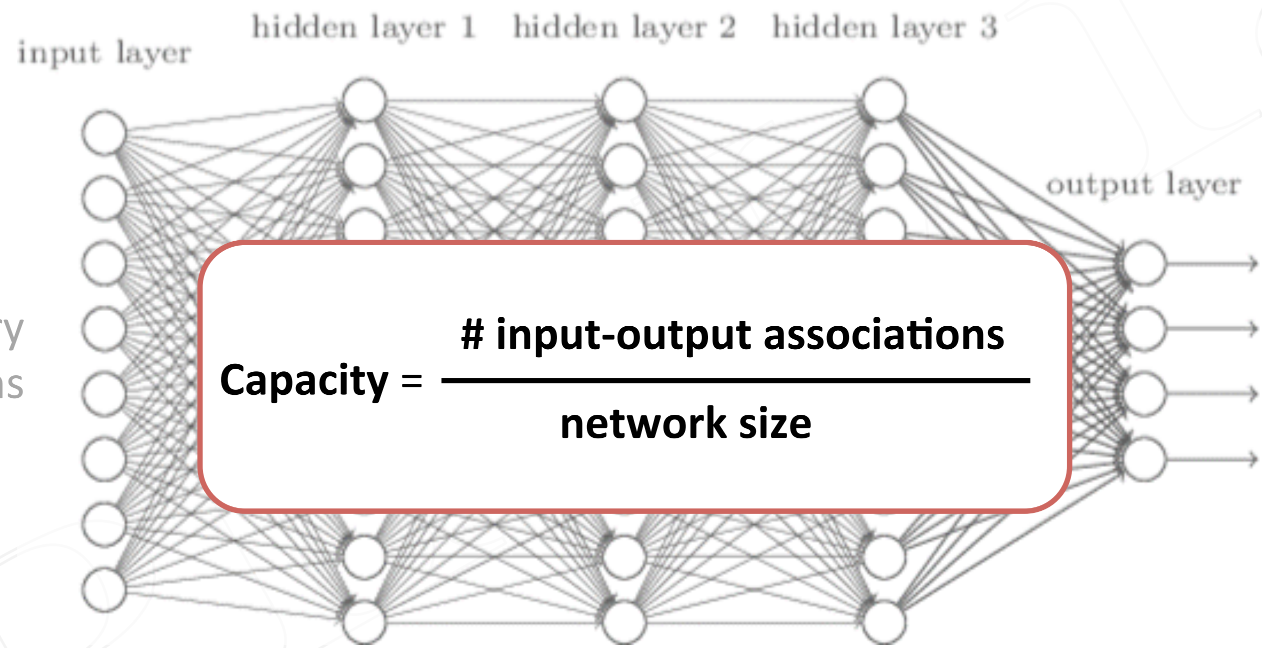
random binary
input patterns



random binary
output labels

Limits of computations in deep nets?

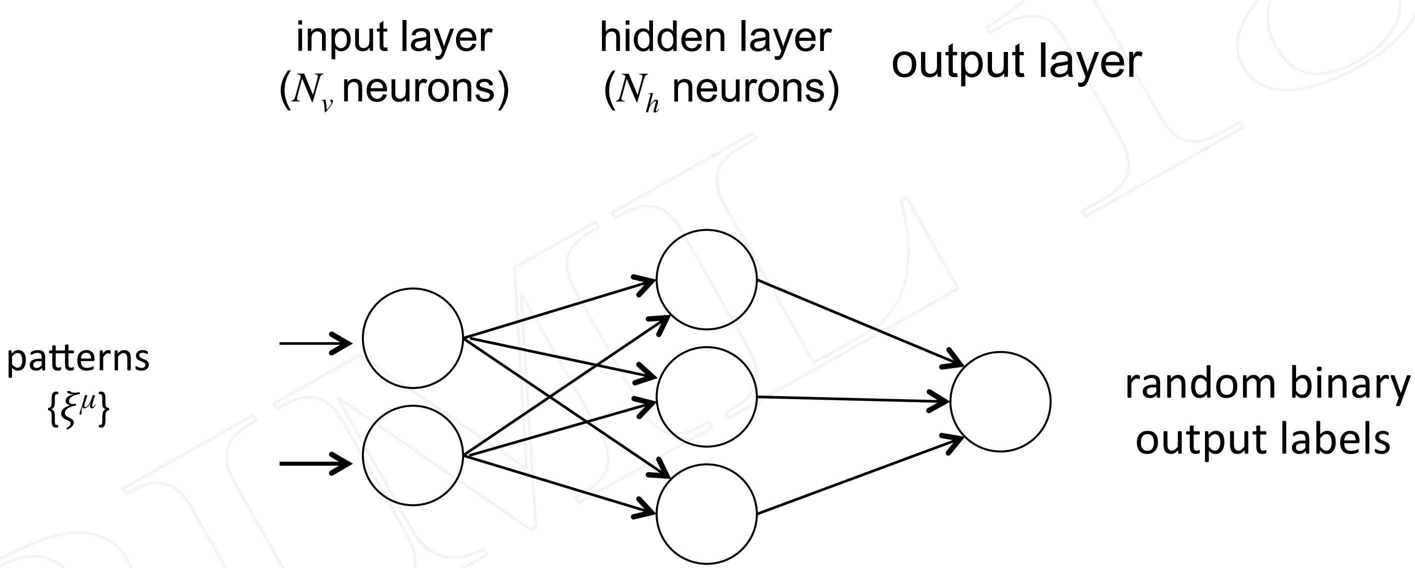
random binary
input patterns



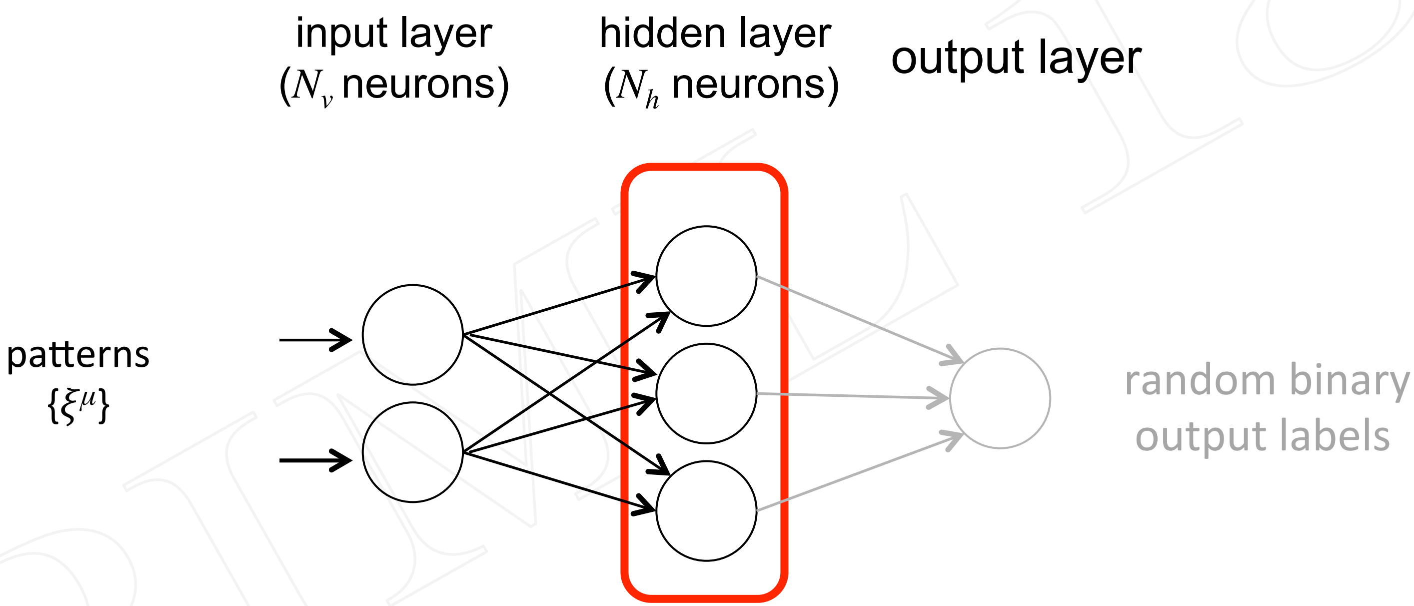
random binary
output labels

$$\text{Capacity} = \frac{\text{\# input-output associations}}{\text{network size}}$$

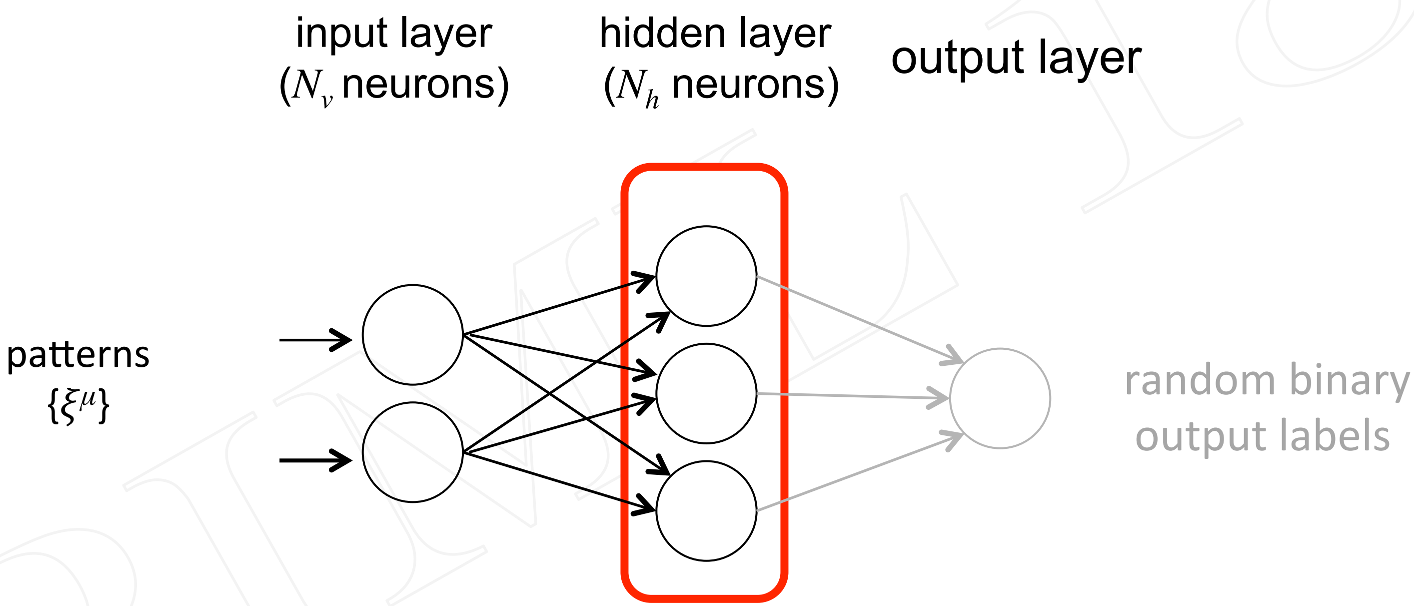
Capacity of multi-layer perceptron (MLP) with one hidden layer?



Investigating internal representation



Expansive (or *overcomplete*) representation

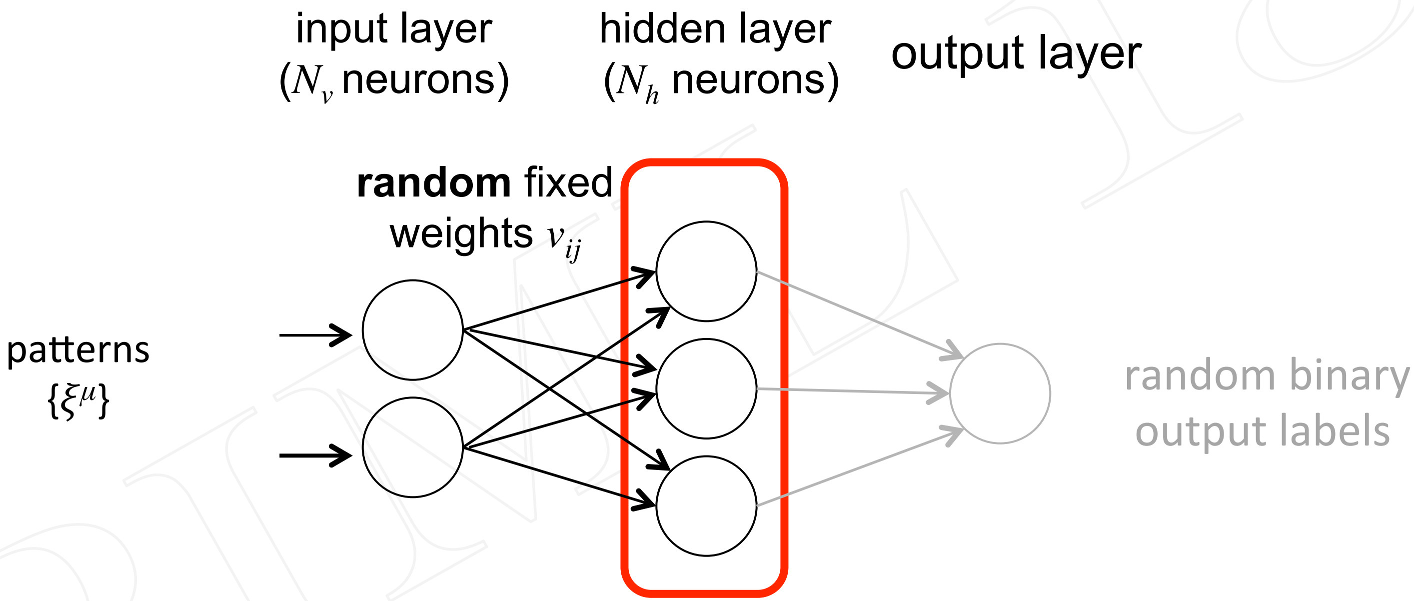


$$\Lambda = \frac{N_h}{N_v} > 1$$

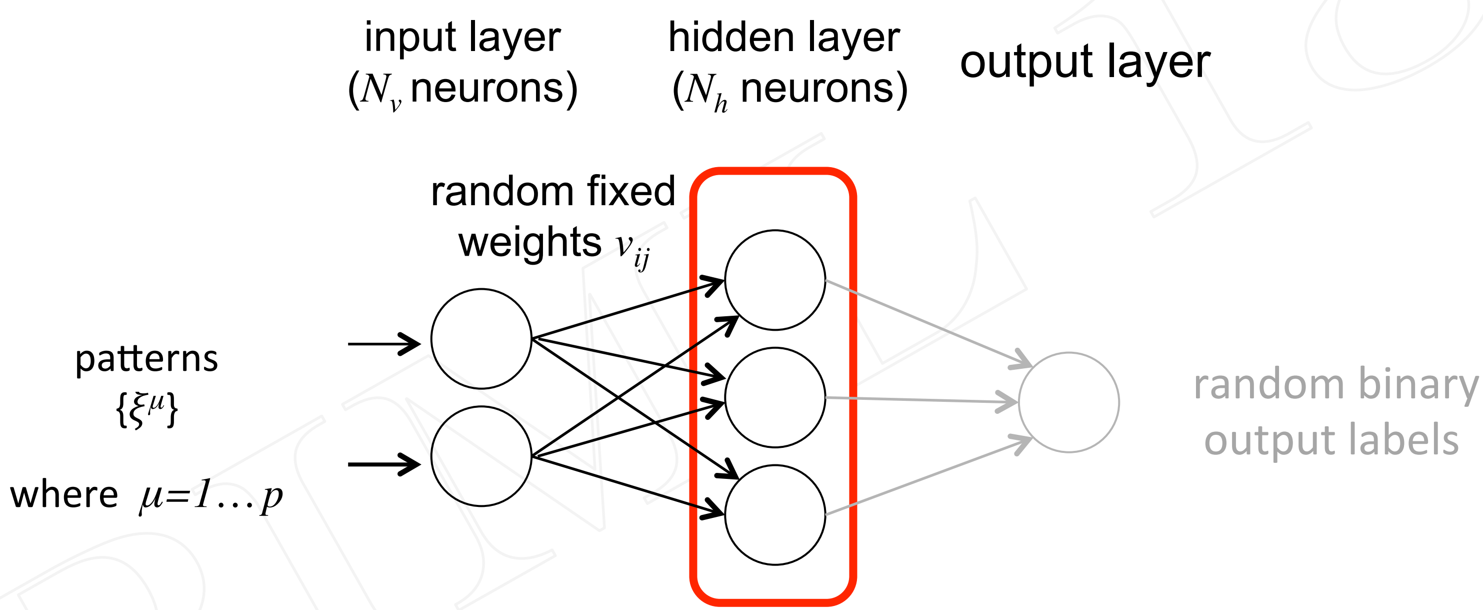
Expansive (or *overcomplete*) representation

Nervous system	Expansion ratio (Λ)
Entorhinal cortex → Dentate gyrus	~ 10
LGN → V1	~ 25
in fly olfactory system: antennal lobe → mushroom body	~ 40
in the cerebellum, mossy fibers → granule cells	$100 \sim 200$
rodent's olfactory bulb → piriform cortex	$\sim 10^3$

Expansive representation with **random** weights

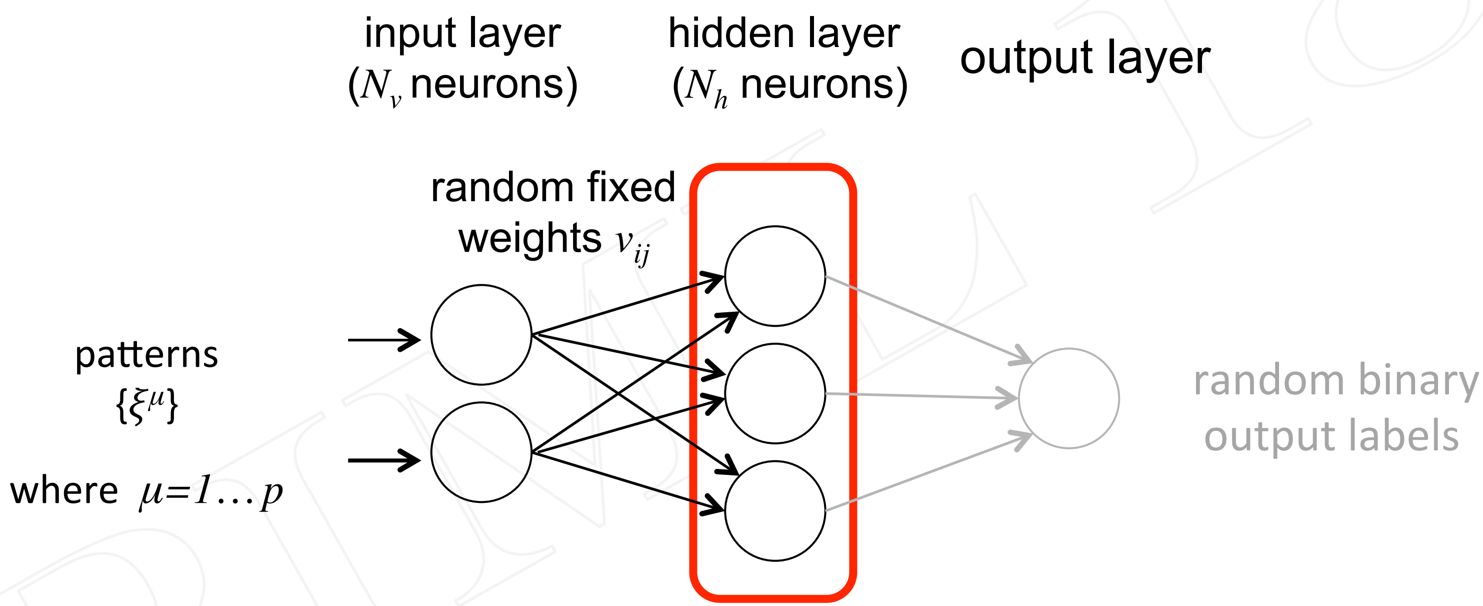


Expansive representation with random weights



How good is random-weight expansive representations?

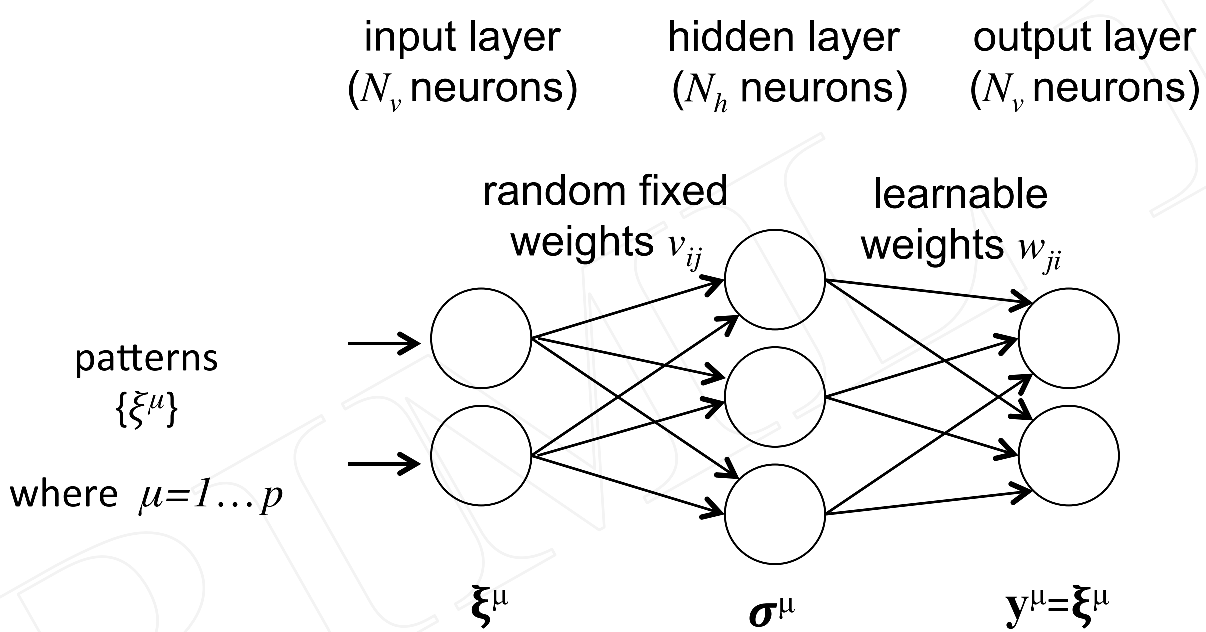
Expansive representation with random weights



How good is ra
expansive repr

How many of input patterns can be reconstructed from this representation?

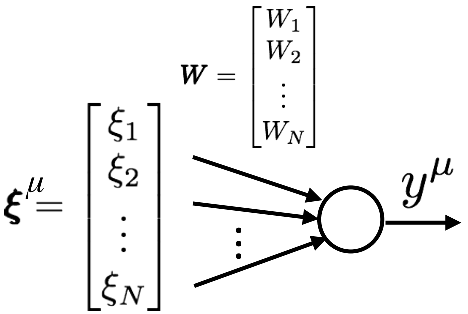
Expansive **autoencoder** with random weights



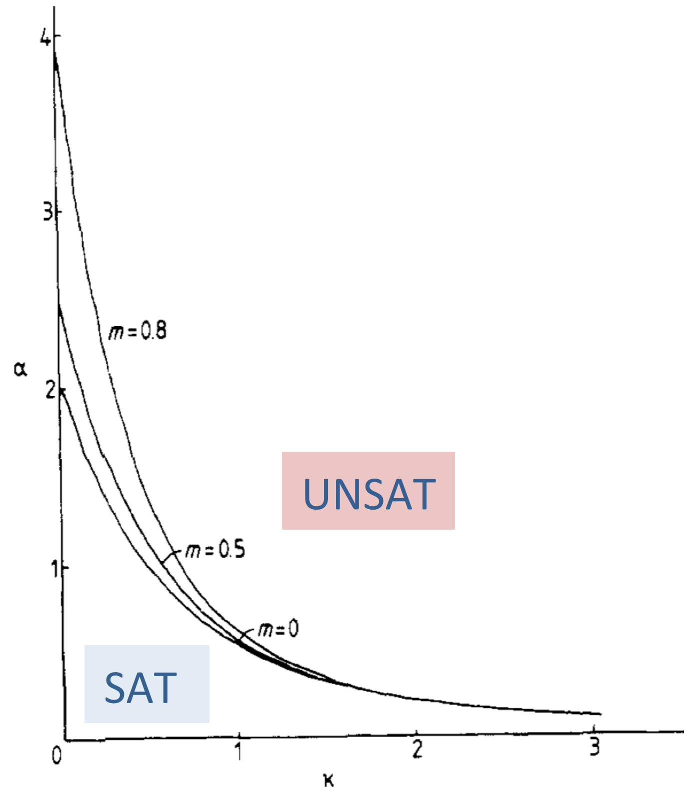
Focus of this talk:

Optimal capacity of expansive autoencoder?

Critical capacity of the perceptron: Gardner approach



$$y^{\mu} = \mathrm{sgn}\left(\frac{\mathbf{w}\cdot\boldsymbol{\xi}^{\mu}}{\sqrt{N}} - \theta\right)$$



Gardner E. (1988) J. Phys. A: Math. Gen.

- Find $\mathbf{w}$ to store a set of random ensemble of input-output pairs $\{ (\boldsymbol{\xi}^{\mu}, d^{\mu}) \}$, $\mu=1\ldots p$ with a robustness κ
- Storage criteria: $\forall \mu, i: \; d^{\mu}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i \xi_i^{\mu} - \theta\right) > \kappa$

Gardner volume ($N \longrightarrow \infty$):

$$\Omega = \int_{\|\mathbf{w}\|^2=N} d^N\mathbf{w} \prod_{\mu=1}^{p} \Theta\left(d^{\mu}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i \xi_i^{\mu} - \theta\right) - \kappa\right)$$

$\langle\!\langle \log(\Omega) \rangle\!\rangle$ averaging over the (quenched) distribution of the patterns

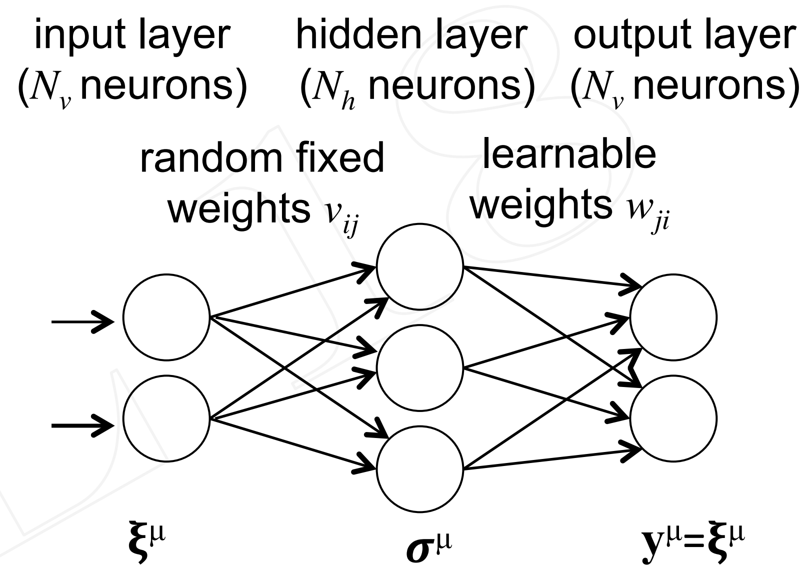
$$\langle\!\langle \log \Omega \rangle\!\rangle = \lim_{n\to 0} \frac{\langle\!\langle \Omega^n \rangle\!\rangle - 1}{n}, \quad \text{critical capacity: } \alpha_c = \lim_{N\to\infty} \frac{p_c}{N}$$

- Perceptron learning rule ($\kappa=0$): $\Delta w_i = \eta \xi_i \,(d^{\mu} - y^{\mu})$

A simplified expansive (over-complete) autoencoder

- Binary neurons (± 1)
- Expansion ratio: $\Lambda = N_h / N_v$
- Fixed encoding weights v_{ij} are randomly sampled from $\mathcal{N}(0, 1)$
- Goal: reconstruct random patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1,\dots,p}$ by learning the decoding weights w_{ji}
- Perceptron learning rule: $\Delta w_{ji} = \eta(\xi_j^\mu - y_j^\mu)\sigma_i^\mu$
- Dense coding level for patterns: $P(\xi_i = +1) = 0.5$
- Coding level of hidden layer: $P(\sigma_i = +1) = f$
- Maximal/critical storage Capacity:
  $\alpha_{max} = p_{max} / N_h$ as $N_h \longrightarrow \infty$
- Storage criteria: meeting fixed-point equations

$$\forall j, \mu: \quad \xi_j^\mu = \operatorname{sgn}\left(\sum_i w_{ji} \operatorname{sgn}\left(\sum_l v_{il}\xi_l^\mu\right)\right)$$



Dynamics:

$$\sigma_i^\mu = \operatorname{sgn}\left(\sum_{j=1}^{N_v} v_{ij}\xi_j^\mu - \theta\right)$$

$$y_j^\mu = \operatorname{sgn}\left(\sum_{i=1}^{N_h} w_{ji}\sigma_i^\mu\right)$$

Alemi A., Abbara A, 2017 *ArXiv*

A mean-field approximation (MFA)

$$\xi_j^\mu = \mathrm{sgn}\left(\sum_{i=1}^{N_h} w_{ji}\mathrm{sgn}\left(\sum_{l=1}^{N_v} v_{il}\xi_l^\mu\right)\right)$$

Approximated by a *quenched* random variable with a Gaussian distribution : $\mathcal{N}(0, N_v)$

$$= \mathrm{sgn}\left(\sum_{i=1}^{N_h} w_{ji}\mathrm{sgn}\left(\boxed{\sum_{l\neq j;l=1}^{N_v} v_{il}\xi_l^\mu} + v_{ij}\xi_j^\mu\right)\right)$$

$$= \mathrm{sgn}\left(\sum_{i=1}^{N_h} w_{ji}\mathrm{sgn}\left(z_i^\mu + v_{ij}\xi_j^\mu\right)\right)$$

$$\mathrm{P}\left(\sigma_i^\mu \middle| \xi_j^\mu\right) \simeq \frac{1}{2} + \sigma_i^\mu \xi_j^\mu \frac{v_{ij}}{\sqrt{2\pi N_v}}$$

$$\left(\sigma_i^\mu \perp \sigma_k^\mu\right) \middle| \, \xi_j^\mu \quad \Rightarrow \quad \mathrm{P}(\boldsymbol{\sigma}^\mu | \xi_j^\mu) = \prod_i \mathrm{P}(\sigma_i^\mu | \xi_j^\mu)$$

A mean-field approximation (MFA)

Probability distribution

$$\mathrm{P}\left(\sigma_i^\mu \middle| \xi_j^\mu\right) \simeq \frac{1}{2} + \sigma_i^\mu \xi_j^\mu \frac{v_{ij}}{\sqrt{2\pi N_v}}$$

$$\left(\sigma_i^\mu \perp \sigma_k^\mu\right) \middle| \, \xi_j^\mu \quad \Longrightarrow \quad \mathrm{P}(\boldsymbol{\sigma}^\mu | \xi_j^\mu) = \prod_i \mathrm{P}(\sigma_i^\mu | \xi_j^\mu)$$

Mean-Field Approximation (MFA) for the *replica* calculation

Gardner volume:

$$\Omega = \int_{\|\mathbf{w}\|^2=N_h} d^{N_h}\mathbf{w} \prod_{\mu=1}^{p} \Theta\Big(\xi_j^\mu \frac{1}{\sqrt{N_h}} \sum_{i=1}^{N_h} w_{ji}\sigma_i^\mu - \kappa\Big)$$

Replica theory: compute $\langle\!\langle \log(\Omega) \rangle\!\rangle$ over distribution of patterns using the replica method

$$\langle\!\langle \log \Omega \rangle\!\rangle = \lim_{n\to 0} \frac{\langle\!\langle \Omega^n \rangle\!\rangle - 1}{n}$$

Probability distribution

$$\mathrm{P}\big(\sigma_i^\mu \big| \xi_j^\mu\big) \simeq \frac{1}{2} + \sigma_i^\mu \xi_j^\mu \frac{v_{ij}}{\sqrt{2\pi N_v}}$$

$$\big(\sigma_i^\mu \perp \sigma_k^\mu\big) \,\big|\, \xi_j^\mu \quad \Longrightarrow \quad \mathrm{P}(\boldsymbol{\sigma}^\mu | \xi_j^\mu) = \prod_i \mathrm{P}(\sigma_i^\mu | \xi_j^\mu)$$

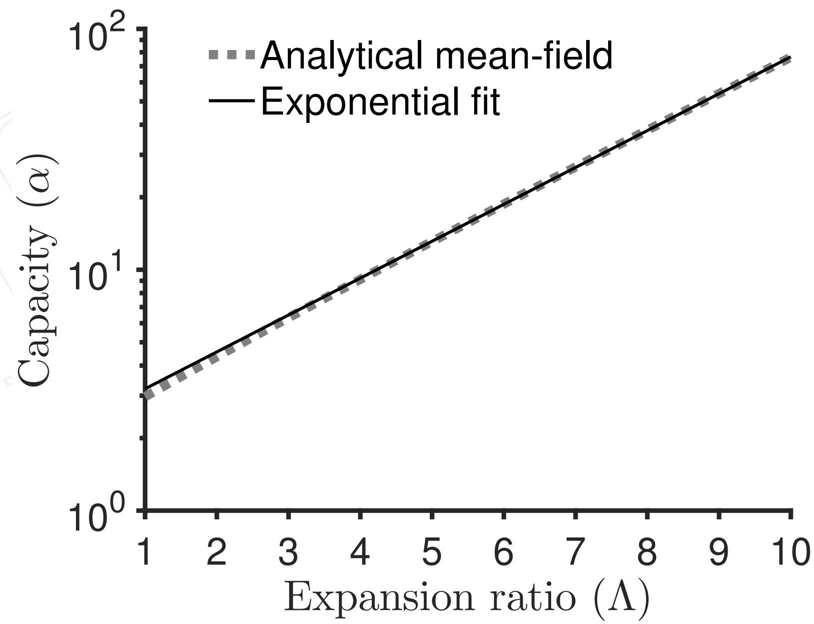
Exponential critical capacity in the MFA

Solutions (replica-symmetric, locally stable):

$$\boxed{\begin{aligned}\alpha_c^{\mathrm{MFA}} \int_{M\sqrt{2\Lambda/\pi}-\kappa}^{\infty} Dt\left(\kappa + t - \sqrt{2\Lambda/\pi}\right)^2 &= 1 - M^2 \\ \alpha_c^{\mathrm{MFA}} \int_{M\sqrt{2\Lambda/\pi}-\kappa}^{\infty} Dt\left(\kappa + t - \sqrt{2\Lambda/\pi}\right)\sqrt{2\Lambda/\pi} &= M,\end{aligned}}$$

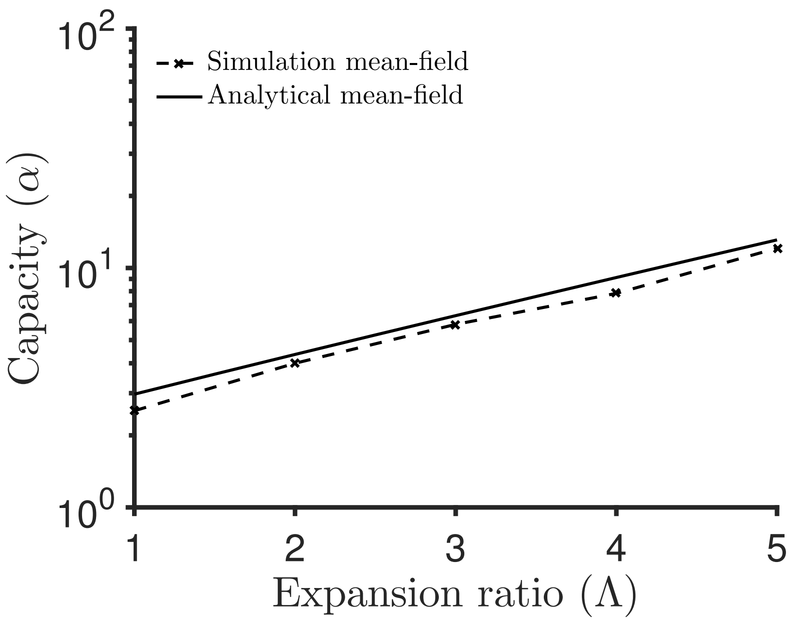
where $M = \sum_i \frac{v_i w_i}{N_h},\ \Lambda = \frac{N_h}{N_v}$

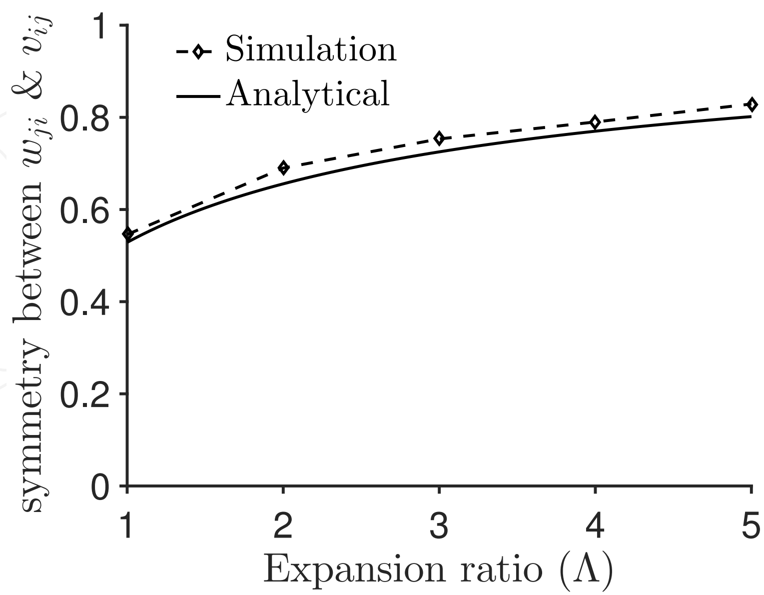
$$Dt \equiv \frac{dt}{\sqrt{2\pi}} \mathrm{e}^{\frac{-t^2}{2}}$$



Alemi A., Abbara A, 2017 *ArXiv*

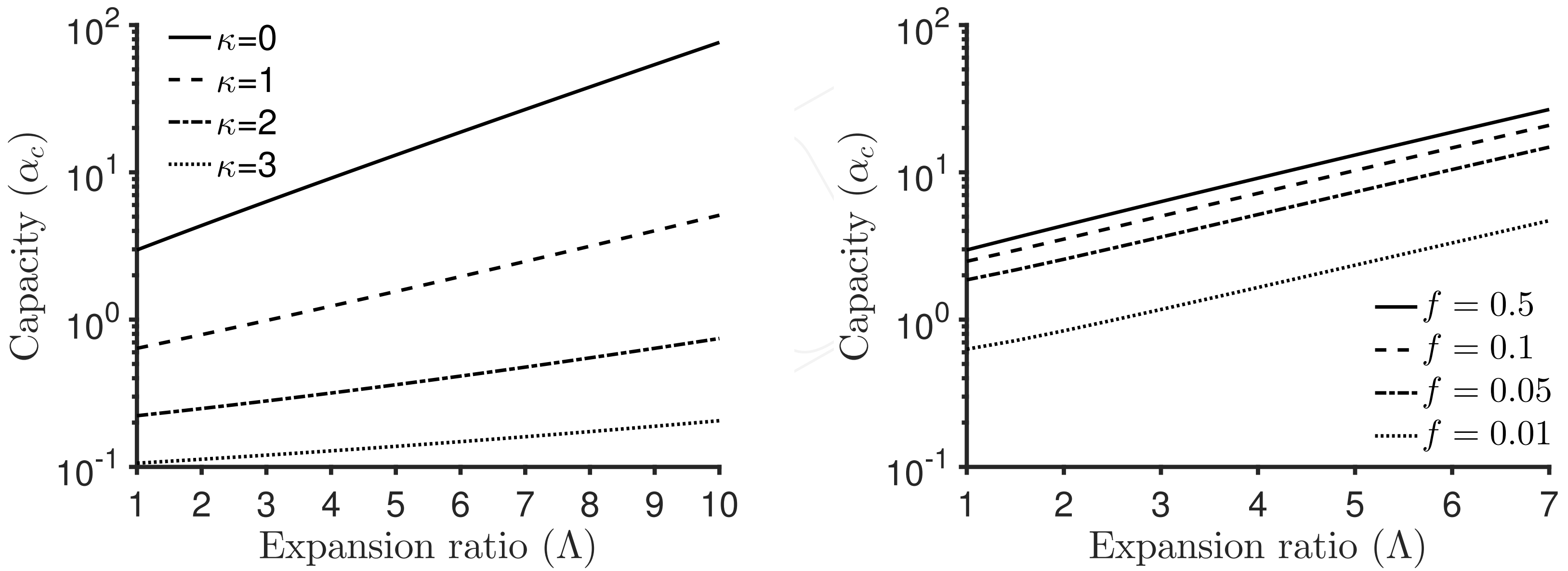
Comparison with simulation (perceptron learning rule)





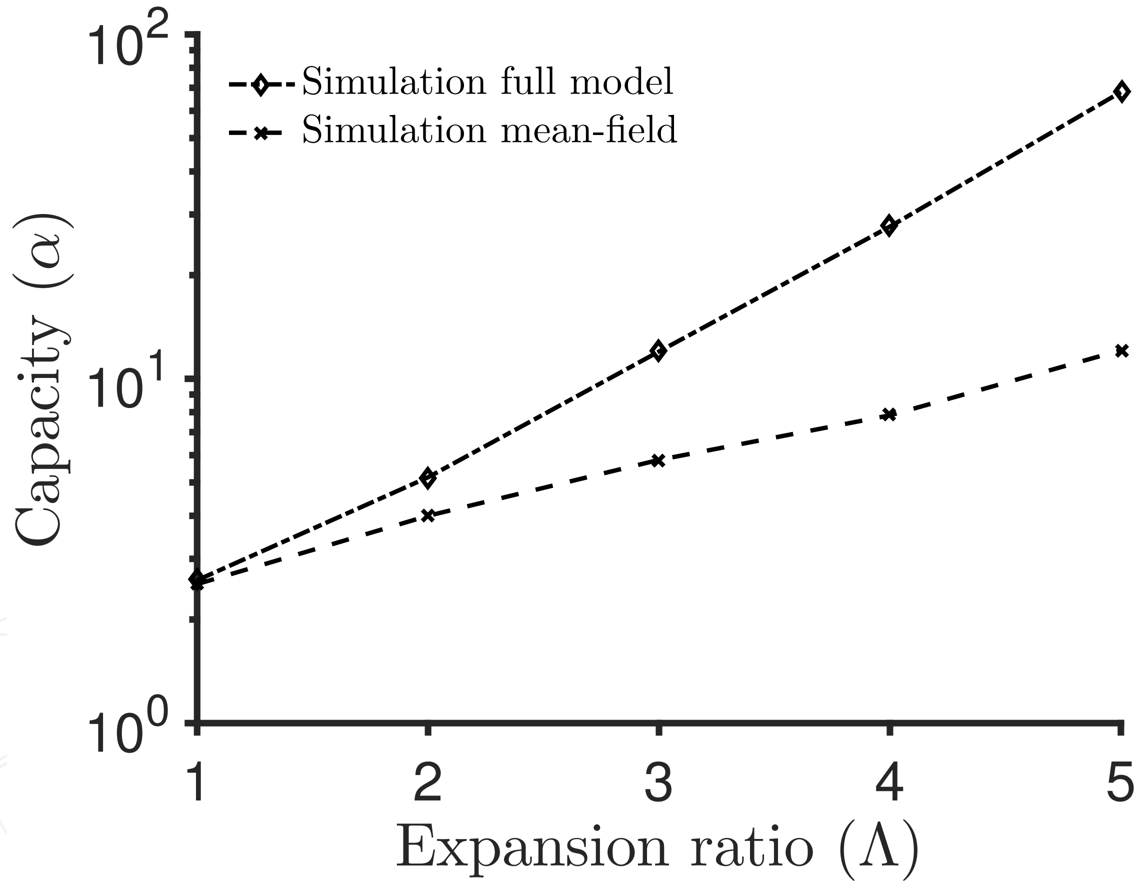
Alemi A., Abbara A, 2017 *ArXiv*

Effect of robustness of output, sparseness in the hidden layer



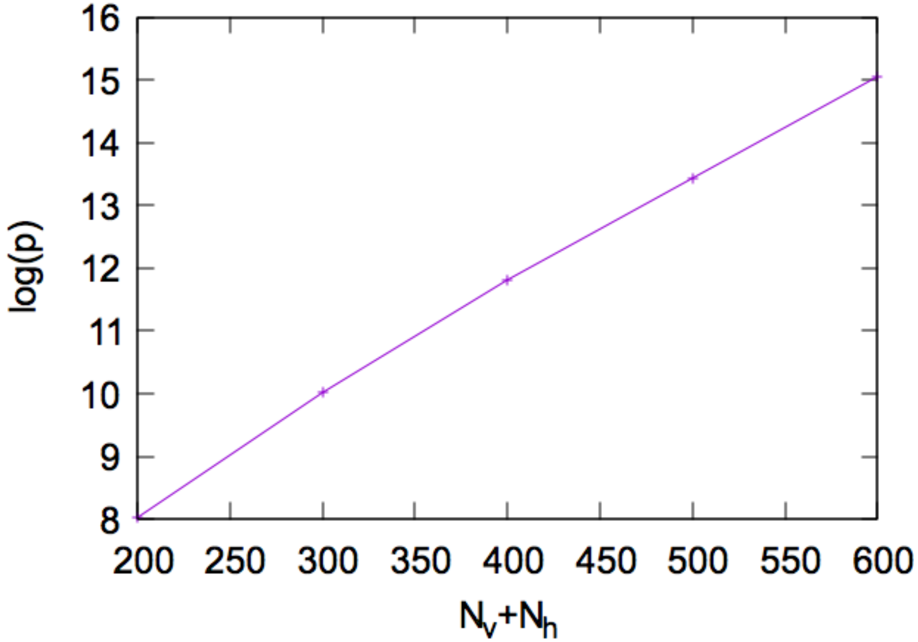
Alemi A., Abbara A, 2017 *ArXiv*

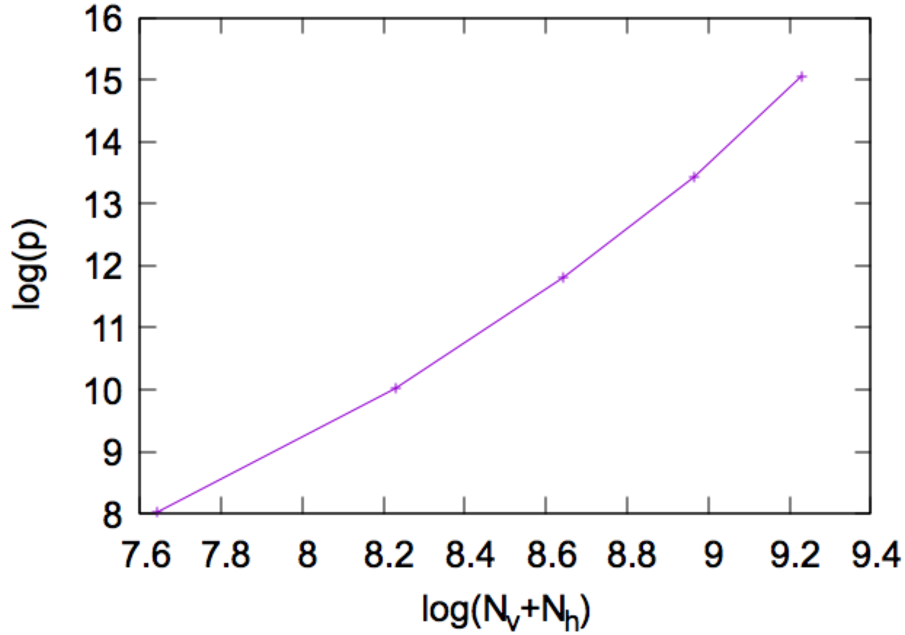
Comparison with the full-model



Alemi A., Abbara A, 2017 *ArXiv*

Log of number of patterns vs. total number of neurons





Discussion

- Expansive representation has important computational implications.
- *Exponential* scaling of the capacity with the expansion ratio in our autoencoder
  - Expansion makes up for the loss of information due to binarization
- In very good agreement with results of simulations using an online learning rule
- Future directions
  - Robustness to noise by training encoding weights
  - Computing the capacity of the full autoencoder model
  - Extension to the capacity of MLP and deep nets

Acknowledgement

Alia Abbara

Discussion with many colleagues …

Question?